

# Goodness of fit statistics for sparse contingency tables

Audrey Finkler

## Abstract

Statistical data is often analyzed as a contingency table, sometimes with empty cells called zeros. Such sparse tables can be due to scarce observations classified in numerous categories, as for example in genetic association studies. Thus, classical independence tests involving Pearson's chi-square statistic  $Q$  or Kullback's minimum discrimination information statistic  $G$  cannot be applied because some of the expected frequencies are too small. More generally, we consider goodness of fit tests with composite hypotheses for sparse multinomial vectors and suggest simple corrections for  $Q$  and  $G$  that improve and generalize known procedures such as Ku's. We show that the corrected statistics share the same asymptotic distribution as the initial statistics. We produce Monte Carlo estimations for the type I and type II errors on a toy example. Finally, we apply the corrected statistics to independence tests on epidemiological and ecological data.

## 1 Introduction and notations

Physical, sociological or biological surveys often lead to data presented as contingency tables. These surveys aim at studying relationships such as total, mutual, partial or conditional independence between several characters in a population. Table notations are very useful to formulate the test and make the corresponding hypotheses explicit, but can be cumbersome when the number of characters exceeds three. For theoretical results, we therefore use vector notations instead, and we reformulate the independence test as a multinomial goodness of fit test.

### 1.1 Goodness of fit tests

Let  $p = (p_1, \dots, p_R)$  be a probability vector of dimension  $R$  where  $R \geq 2$  is the total number of cross-classifying categories. Let  $n$  be the sample size and  $x = (n_1, \dots, n_R)$  the vector of observed frequencies, realization of  $X = (N_1, \dots, N_R)$  with distribution  $\mathcal{M}(n; p)$ , multinomial distribution with parameters  $n$  and  $p$ . We denote by  $p^0$  the probability vector under the null

hypothesis and consider the following test :

$$\mathcal{H}_0 : p = p^0 \quad \text{against} \quad \mathcal{H}_1 : p \neq p^0. \quad (1.1)$$

Popular goodness of fit statistics include Pearson's chi-square statistic  $Q$  and Kullback's minimum discrimination information statistic  $G$ , defined in [7, 10]. For two probability distributions  $p$  and  $p'$ , these are written :

$$Q_p(p') = n \sum_{r=1}^R \frac{(p'_r - p_r)^2}{p_r}, \quad (1.2)$$

and

$$G_p(p') = 2n \sum_{r=1}^R p'_r \ln \frac{p'_r}{p_r}. \quad (1.3)$$

They belong to the power divergence statistics family  $\{RC^\lambda, \lambda \in \mathbb{R}\}$  defined by Read and Cressie in [5], respectively for  $\lambda = 1$  and  $\lambda \rightarrow 0$ , where :

$$RC_p^\lambda(p') = \frac{2n}{\lambda(\lambda + 1)} \sum_{r=1}^R p'_r \left[ \left( \frac{p'_r}{p_r} \right)^\lambda - 1 \right]. \quad (1.4)$$

For a review on goodness of fit testing methods and statistics, see [3].

The vector  $p^0$  is not always completely specified. We assume that  $p^0$  is a known function of an unknown parameter  $\theta$  of  $\Theta \subseteq \mathbb{R}^s$  with  $s < R - 1$ , and denote  $p^0 = p^0(\theta)$  with  $\theta = (\theta_1, \dots, \theta_s)$ . We suppose that the functions  $\theta \mapsto p^0(\theta)$  we consider here are bijective and estimate  $p^0(\theta)$  by  $p^{*0} = p^0(\theta^*)$  where  $\theta^*$  is the maximum likelihood estimator of  $\theta$ . As the probability  $p$  is generally unknown, we also estimate the  $p_r$  for  $r$  in  $\{1, \dots, R\}$  by their maximum likelihood estimators  $p_r^* = n_r/n$ . Throughout this paper, underlying indexes  $n$  are omitted for simplicity of notation, and we consider the statistics  $Q_{p^{*0}}(p^*)$  and  $G_{p^{*0}}(p^*)$ .

Read and Cressie show that under Birch's regularity conditions, see [2], the statistics  $RC^\lambda$  are asymptotically equivalent and share a common chi-square limit distribution:

**Theorem 1.**

$$\forall \lambda \in \mathbb{R}, \quad \lim_{n \rightarrow +\infty} \mathcal{L}_{\mathcal{H}_0}(RC_{p^{*0}}^\lambda(p^*)) = \chi_{R-s-1}^2. \quad (1.5)$$

This implies the following result :

**Corollary 1.**

$$\lim_{n \rightarrow +\infty} \mathcal{L}_{\mathcal{H}_0}(Q_{p^{*0}}(p^*)) = \lim_{n \rightarrow +\infty} \mathcal{L}_{\mathcal{H}_0}(G_{p^{*0}}(p^*)) = \chi_{R-s-1}^2. \quad (1.6)$$

This asymptotic result is a consequence of the Central Limit Theorem. Classical empirical limitations include that the sample size  $n$  must be over 30 and that all expected frequencies must be over 5. Usually  $G_{p^{*0}}(p^*)$  is preferred to  $Q_{p^{*0}}(p^*)$  because it is less sensitive to small cell frequencies. Read and Cressie recommend the use of  $RC_{p^{*0}}^{2/3}(p^*)$  instead for  $n \geq 10$  and a minimum expected frequency over 1. Sometimes, even the lower bound 0.5 is accepted for the expected frequencies. A review on these conditions can be found in [4].

## 1.2 Sparse tables

When there are too few subjects in the study or when the classifying categories are too numerous, the table comprises one or several empty cells called random zeros. The table is then called sparse and it is likely that at least one cell has an expected frequency below 0.5. Random zeros would not appear if the sample was of sufficient size. Structural zeros, corresponding to cells with an expected probability of zero, are not considered here and should be suppressed. We therefore assume that the following condition is satisfied :

$$p_r^{*0} \neq 0, r \in \{1, \dots, R\}. \quad (1.7)$$

Sparse tables can nonetheless be tested for independence, for example by regrouping cells so that the condition on the expected frequencies is satisfied. However this procedure is not always data relevant. Fisher's exact test given in [6] applies without restrictions, except that it becomes numerically unmanageable when the table dimension grows. This leads to the use of Monte Carlo simulation methods as explained in [1].

The approach we propose here consists in correcting the historical statistics  $Q_{p^{*0}}(p^*)$  and  $G_{p^{*0}}(p^*)$  according to the number of zero cells, by generalizing and improving a method designed by Ku in [9].

## 2 Corrections for Pearson's and Kullback's statistics

Let  $C$  be the random variable giving the number of zeros in the vector or the contingency table, and  $c$  a realization of  $C$  with  $R - c \geq 1$ . For simplicity, we assume that  $n_1 = n_2 = \dots = n_c = 0$  and that  $n_j \geq 1$  for all  $j$  in  $\{c + 1, \dots, R\}$ . The maximum likelihood estimator  $p^* = (0, \dots, 0, n_{c+1}/n, \dots, n_R/n)$  underestimates the  $p_i$  for  $i$  in  $\{1, \dots, c\}$  and overestimates the  $p_j$  for  $j$  in  $\{c + 1, \dots, R\}$ . Its use when  $c \neq 0$  thus has consequences on the statistics  $Q_{p^{*0}}(p^*)$  and  $G_{p^{*0}}(p^*)$ .

## 2.1 Ku's correction for one zero

Ku argues in [9] that  $G_{p^*0}(p^*)$  tends to inflate with respect to  $Q_{p^*0}(p^*)$  when  $C$  grows. He then proposes to subtract 1 from  $G_{p^*0}(p^*)$  for each zero, that is  $c$  in total. He proves the asymptotic equivalence between  $G_{p^*0}(p^*)$  and its corrected version only for  $c = 1$ . I will explain why his reasoning is inconsistent. First, he considers a new statistic which is not a member of the power divergence family. Moreover, he uses the straightforward Lemma 1 to deduce the approximation:

$$2n_r \ln \left( \frac{n_r}{np_r^{*0}} \right) \simeq \frac{n_r^2 - (np_r^{*0})^2}{np_r^{*0}}, \quad (2.1)$$

for  $a = n_r/n$ ,  $b = p_r^{*0}$  and  $n_r = 0$ , despite the fact that  $a = 0$  and  $1/a$  is not bounded.

**Lemma 1.** *For each  $a, b > 0$  such that  $a < 2b$ ,  $b < 2a$  and the quantities  $1/a$  and  $1/b$  are bounded, we have :*

$$\ln \left( \frac{a}{b} \right) = \frac{a^2 - b^2}{2ab} + o(a - b).$$

The expression on the left in (2.1) is null whereas the one on the right is negative. The sum over  $r$  of the left-hand side gives  $G_{p^*0}(p^*)$  and we recognize  $Q_{p^*0}(p^*)$  in the sum of the right-hand side. However, unlike Ku seems to think, zero and non-zero cells tend to compensate for each other and the approximation (2.1) can not be extended to the corresponding sum. There is therefore no behavior of  $G$  and  $Q$  that we can deduce from this. Finally, he illustrates this asymptotic result on a small sample of size  $n = 10$ .

We propose new corrections for both statistics, based on Ku's correction and a likelihood inequality that we present in the next subsection.

## 2.2 Likelihood inequality

Let us consider the following inequality coming from a likelihood reasoning : the sample vector we observe can be thought of more likely to happen than any other possible vector, since it is the one we actually observed. With exactly  $c$  zeros and  $R - c$  non-zero cells observed, so for all  $m \leq c$  and  $n'_j$  such that  $n'_j \leq n_j$ ,  $\forall j \in \{c+1, \dots, R\}$  and  $n = \sum_{i=1}^m n'_i + \sum_{j=c+1}^R n'_j$  :

$$\begin{aligned} \mathbb{P}(N_1 = 0, \dots, N_c = 0, N_{c+1} = n_{c+1}, \dots, N_R = n_R) &\geq \\ \mathbb{P}(N_1 = n'_1, \dots, N_m = n'_m, N_{m+1} = \dots = N_c = 0, N_{c+1} = n'_{c+1}, \dots, N_R = n'_R). \end{aligned} \quad (2.2)$$

We give in Proposition 1 a sufficient condition on the  $p_r$  for (2.2) to be satisfied, and prove this statement in Appendix A.

**Proposition 1.** *The inequality (2.2) is satisfied under the following assumption :*

$$p_i \leq \frac{p_j}{n}, \quad \forall i \in \{1, \dots, c\}, \quad \forall j \in \{c+1, \dots, R\}. \quad (2.3)$$

### 2.3 Corrections

We propose an estimator  $\hat{p}$  for  $p$ , different from the maximum likelihood estimator, with the following form :

$$\begin{cases} \hat{p}_i &= a, & \forall i \in \{1, \dots, c\}, \\ \hat{p}_j &= \frac{n_j}{n^b} - d, & \forall j \in \{c+1, \dots, R\}, \end{cases} \quad (2.4)$$

where  $a = a_n$ ,  $b = b_n$  and  $d = d_n$  are random variables depending on  $n$  designed to compensate for the under- and overestimations due to  $p^*$ . We thus take them positive with  $0 < b < 1$ , such that  $b$  inflates the modified maximum likelihood estimator  $n_j/n^b$  and such that  $d$  controls the related rise.

This new probability vector allows us to define corrected statistics  $Q_{p^*0}(\hat{p})$  and  $G_{p^*0}(\hat{p})$ , provided that the parameters  $a$ ,  $b$  and  $d$  satisfy several conditions. Forcing the summation of the  $\hat{p}_r$  to 1 implies that  $(R - c)d = ac + n^{1-b} - 1$ , and thus allows us to define  $Q_{p^*0}(\hat{p}^{ab})$  and  $G_{p^*0}(\hat{p}^{ab})$  with  $\hat{p}^{ab}$  such that :

**Definition 1.**

$$\begin{cases} \hat{p}_i^{ab} &= a, & \forall i \in \{1, \dots, c\}, \\ \hat{p}_j^{ab} &= \frac{n_j}{n^b} - \frac{ac + n^{1-b} - 1}{R - c}, & \forall j \in \{c+1, \dots, R\}. \end{cases} \quad (2.5)$$

Note that for  $c = 0$ , we fix  $a = 0$  and  $b = 1$ .

Before considering the other conditions on  $\hat{p}^{ab}$ , let us first give some notations. Let  $\underline{n}$  be  $\min_{j \in \{c+1, \dots, R\}} \{n_j\}$  and  $\overline{n}$  be  $\max_{j \in \{c+1, \dots, R\}} \{n_j\}$ . Let  $\underline{\underline{n}}$  stand for  $n - \underline{n}(R - c)$  and  $\overline{\overline{n}}$  for  $\overline{n}(R - c) - n$ . We dismiss the uniformly distributed case where :

$$\underline{n} = \overline{n} = n_j = \frac{n}{R - c}, \quad \forall j \in \{c+1, \dots, R\}, \quad (2.6)$$

thus guaranteeing that  $\underline{\underline{n}}$  and  $\overline{\overline{n}}$  are positive. Let

$$b_{\min} = \max \left( 0, \frac{\ln(\overline{\overline{n}}/(R - 1))}{\ln(n)}, \frac{\ln(\underline{\underline{n}})}{\ln(n)}, \frac{\ln(\overline{n} - \underline{n})}{\ln(n)} \right), \quad (2.7)$$

$$a_{\min}(b) = \max \left( 0, \frac{(\overline{n} - n^b)(R - c) + n^b}{cn^b} \right), \quad (2.8)$$

and

$$a_{\max}(b) = \min \left( 1, \frac{n^b - n}{cn^b}, \frac{n^b - n}{n^b(n(R-c) + c)} \right), \quad (2.9)$$

Proposition 1 is applied to  $\hat{p}^{ab}$ . Together with inequalities  $0 < \hat{p}_r^{ab} < 1$ , for all  $r$  in  $\{1, \dots, R\}$  it is then equivalent to these conditions on  $a$  and  $b$ :

$$b_{\min} < b < b_{\max} = 1, \quad (2.10)$$

and

$$a_{\min}(b) < a < a_{\max}(b). \quad (2.11)$$

We want to make a practical choice among possible values of  $\hat{p}^{ab}$  ensuring us that it is as far from  $p^*$  as possible. We therefore fix  $b$  in (2.10) quite far from 1 as a convex combination of  $b_{\min}$  and  $b_{\max}$  with an empirical parameter  $h$  equal to 0.1. For this value of  $b$  we choose  $a$  near the upper limit of the interval in (2.11), that is :

$$b = hb_{\max} + (1-h)b_{\min} \quad \text{and} \quad a = a_{\max}(b) - \epsilon, \quad (2.12)$$

where  $\epsilon$  is a small constant designed to eliminate boundary effects.

The final expressions we get for the corrected statistics  $Q^{ab} = Q_{p^{*0}}(\hat{p}^{ab})$  and  $G^{ab} = G_{p^{*0}}(\hat{p}^{ab})$  of  $Q = Q_{p^{*0}}(p^*)$  and  $G = G_{p^{*0}}(p^*)$  are:

$$Q^{ab} = n^{2(1-b)}Q - f(a, b), \quad (2.13)$$

with

$$f(a, b) = n \left( 1 - n^{2(1-b)} + \frac{2n^{1-b}(ac + n^{1-b} - 1)}{R - c} \sum_{j=c+1}^R \frac{n_j}{np_j^{*0}} - a^2 \sum_{i=1}^c \frac{1}{p_i^{*0}} - \left( \frac{ac + n^{1-b} - 1}{R - c} \right)^2 \sum_{j=c+1}^R \frac{1}{p_j^{*0}} \right), \quad (2.14)$$

and

$$G^{ab} = n^{1-b}G - g(a, b), \quad (2.15)$$

where

$$g(a, b) = 2n \left( \frac{ac + n^{1-b} - 1}{R - c} \sum_{j=c+1}^R \ln \left( \frac{n_j(R - c) - n^b(ac + n^{1-b} - 1)}{p_j^{*0}n^b(R - c)} \right) - a \sum_{i=1}^c \ln \left( \frac{a}{p_i^{*0}} \right) - n^{1-b} \sum_{j=c+1}^R \frac{n_j}{n} \ln \left( \frac{n_j(R - c) - n^b(ac + n^{1-b} - 1)}{n_j n^{b-1}(R - c)} \right) \right). \quad (2.16)$$

## 2.4 Convergence

In this paragraph, we show the convergence in distribution of  $Q^{ab}$  and  $G^{ab}$  to a chi-square distribution. Let us first study the parameter  $b$ .

**Property 1.** *Bound  $b_{\min}$  is strictly less than 1. Moreover, if  $\underline{n} = o(n)$  and  $\bar{n} \sim n$  when  $n$  tends to  $+\infty$ , then  $b_{\min}$  and  $b$  tend to 1.*

*Proof.* Inequalities  $\bar{n}(R-c) < nR$ ,  $n-\underline{n}(R-c) < n$  and  $\bar{n}-\underline{n} < n$  show that all three components of the maximum defining  $b_{\min}$  are strictly less than 1. Their order 1 developments as  $n$  tends to  $+\infty$  give their convergence to 1, with the quantities  $R-1$ ,  $R-c$  and  $R-c-1$  bounded.  $\square$

We also study the asymptotic behavior of  $C$ , for once denoted  $C_n$ , and state the following lemma which proof appears in Appendix A.

**Lemma 2.** *The number of zeros  $C_n$  converges almost surely to 0 as  $n$  tends to  $+\infty$ .*

We recalled in section 1 the convergence of Pearson's and Kullback's statistics to a chi-square distribution. In Theorem 2, we state a similar property for the corrected statistics  $Q^{ab}$  and  $G^{ab}$ .

**Theorem 2.** *Under Birch's regularity conditions in [2], the estimator  $\hat{p}^{ab}$  defined by (2.5), (2.10) and (2.11) is such that :*

$$\lim_{n \rightarrow +\infty} \mathcal{L}_{\mathcal{H}_0}(Q^{ab}) = \lim_{n \rightarrow +\infty} \mathcal{L}_{\mathcal{H}_0}(G^{ab}) = \chi_{R-s-1}^2. \quad (2.17)$$

*Proof.* We deduce from Lemma 2 the existence of a set  $\Omega'$  of probability 1 on which we can find a rank  $n_0$  such that  $C_n = 0$  for  $n \geq n_0$ . The variable  $a$  is then set equal to 0 and the variable  $b$  equal to 1. Hence, the estimates  $p^*$  and  $\hat{p}^{ab}$  match, so that  $Q^{ab} = Q$  and  $G^{ab} = G$ . Theorem 1 then completes the proof.  $\square$

The final two sections are dedicated to proving the relevancy of our corrections through simulations and real data analyses.

## 3 Simulations

In this section, simulations confirm the necessity to correct not only  $G$  but also  $Q$ . We compute the statistics  $Q$ ,  $G$ ,  $RC^{2/3} = RC_{p^*0}^{2/3}(p^*)$ ,  $Q^{ab}$  and  $G^{ab}$  on 1 000 vectors of length  $R = 100$  of total frequency  $n = 400$  for each of the four multinomial distributions defined in Table 1 by  $f_1$  to  $f_4$ . Let  $E_r$  denote the expected frequencies for  $r$  in  $\{1, \dots, 100\}$ .

For each set of vectors sharing the same  $c$ , we give the quantile of order  $1 - \alpha = 95\%$  for the five statistics considered. Results are displayed in

Table 1: Multinomial probabilities  $f_1$  to  $f_4$ .

$\mathcal{H}_0$	$ \{r; E_r < 0.5\} $	Probability
$f_1$	20	$(\underbrace{0.0002, \dots, 0.0002}_{20 \text{ times}}, \underbrace{0.01245, \dots, 0.01245}_{80 \text{ times}})$
$f_2$	50	$(\underbrace{0.0002, \dots, 0.0002}_{50 \text{ times}}, \underbrace{0.0198, \dots, 0.0198}_{50 \text{ times}})$
$f_3$	70	$(\underbrace{0.0002, \dots, 0.0002}_{70 \text{ times}}, \underbrace{0.03286667, \dots, 0.03286667}_{30 \text{ times}})$
$f_4$	90	$(\underbrace{0.0002, \dots, 0.0002}_{90 \text{ times}}, \underbrace{0.0982, \dots, 0.0982}_{10 \text{ times}})$

figure 1, as well as a line indicating the chi-square quantile of order  $1 - \alpha = 95\%$  with  $R - 1 = 99$  degrees of freedom  $\chi_{0.95,99}^2 = 123.22$ . Only the center of each graph should be considered because the quantiles for extreme numbers of zeros are computed on too few observations, sometimes only on one or two among the 1 000 simulations in total.

Quantile values of  $Q$  tend to explode as  $c$  grows whereas  $G$  stays quite stable around  $\chi_{0.95,99}^2$ . This behavior is the opposite of the one predicted by Ku. For  $f_1$ , statistics  $Q^{ab}$  and  $G^{ab}$  lead to the rejection of the null hypothesis. For  $f_2$  to  $f_4$  however, their quantiles lie below the critical line and  $\mathcal{H}_0$  is accepted. We thus have compensated for the rise of  $Q$ , and both our corrected statistics are stable.

This analysis is confirmed by the computation of empirical risks of type I for  $\alpha = 0.01, 0.05$  and  $0.1$  as showed in Table 2, and by the power study below. Probabilities  $f_1$  to  $f_4$  are perturbed into  $f'_1$  to  $f'_4$  such that for all  $j$  in  $\{1, 2, 3, 4\}$ :

$$\forall i \in \{1, \dots, 10\}, f'_j(i) = f_j(i) + 1/300, \quad (3.1)$$

$$\forall i \in \{11, \dots, 90\}, f'_j(i) = f_j(i), \quad (3.2)$$

$$\forall i \in \{91, \dots, 100\}, f'_j(i) = f_j(i) - 1/300. \quad (3.3)$$

Vectors are simulated with probabilities  $f_1$  to  $f_4$ , and goodness of fit for  $f'_1$  to  $f'_4$  is tested. The Tables 2 and 3 show that the empirical type I risks are lower for our corrections compared to the classical statistics when  $c$  is quite important, whereas the empirical power is much higher for our corrections when  $c$  is small.

## 4 Applications

We apply the total independence test using the corrected statistics  $Q^{ab}$  and  $G^{ab}$  to two datasets involving two-dimensional tables. For such tables, the



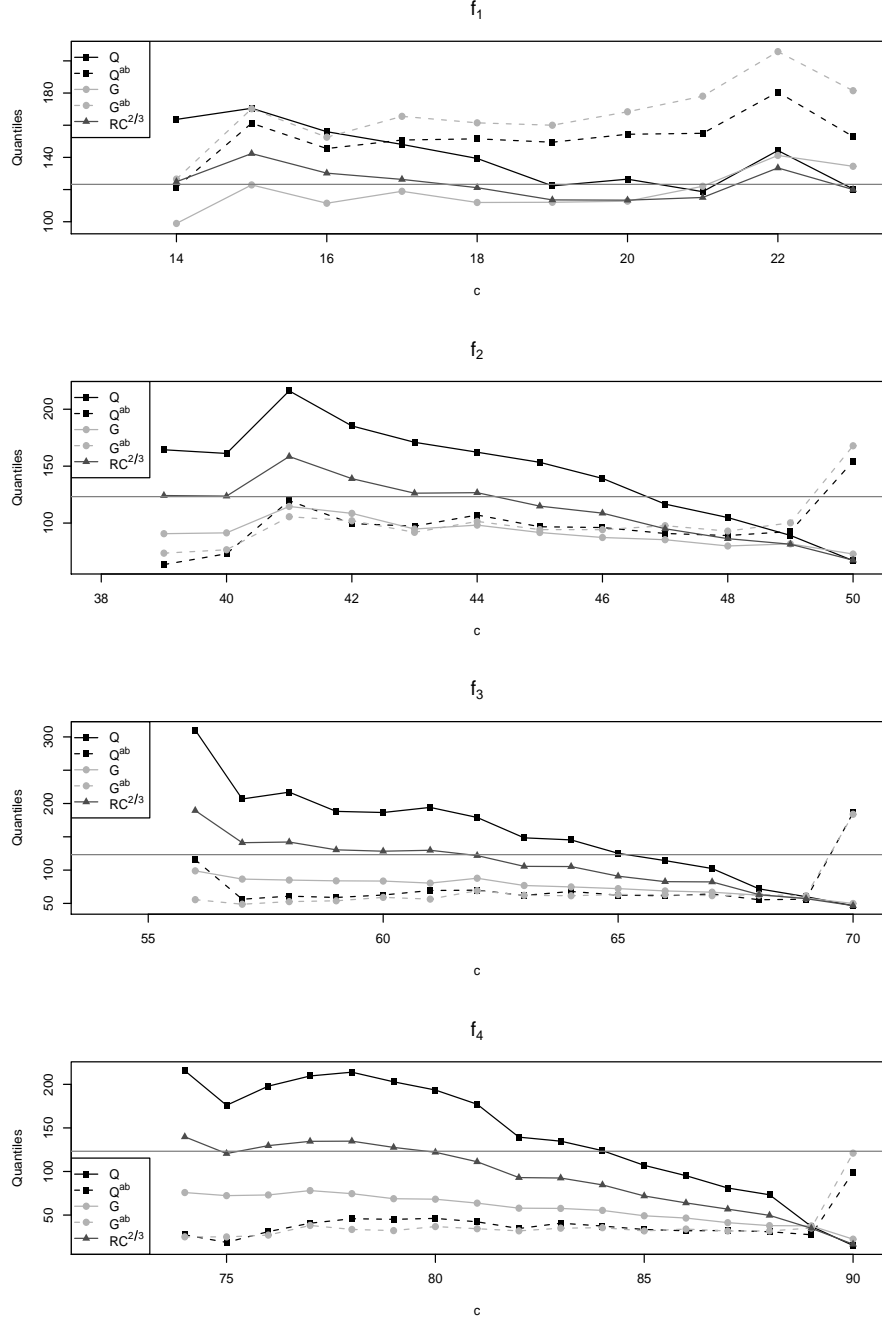


Figure 1: Quantiles of order 0.95 for  $Q$ ,  $Q^{ab}$ ,  $G$ ,  $G^{ab}$  and  $RC^{2/3}$  as functions of  $c$ , under null multinomial probabilities  $f_1$  to  $f_4$ , for 1 000 samples of size  $n = 400$  and  $R = 100$  categories. The line represents the threshold  $\chi^2_{0.95,99}$ .

Table 2: Empirical type I risks for  $Q$ ,  $Q^{ab}$ ,  $G$ ,  $G^{ab}$  and  $RC^{2/3}$ , for 1 000 samples of size  $n = 400$  and  $R = 100$  categories, at levels  $\alpha = 0.01$ ,  $0.05$ ,  $0.1$ , for vector probabilities  $f_1$  to  $f_4$ .

$\alpha$	$\mathcal{H}_0$	$mode(c)$	$Q$	$Q^{ab}$	$G$	$G^{ab}$	$RC^{2/3}$
0.01	$f_1$	19	0.031	0.233	0.003	0.401	0.003
	$f_2$	47	0.030	0.005	0	0	0
	$f_3$	65	0.027	0	0	0	0
	$f_4$	84	0.031	0	0	0	0
0.05	$f_1$	19	0.044	0.352	0.010	0.573	0.017
	$f_2$	47	0.024	0.005	0	0	0
	$f_3$	65	0.069	0	0	0	0.006
	$f_4$	84	0.052	0	0	0	0
0.1	$f_1$	19	0.130	0.479	0.033	0.674	0.039
	$f_2$	47	0.072	0.005	0	0.005	0
	$f_3$	65	0.137	0	0	0	0
	$f_4$	84	0.098	0	0	0	0.006

Table 3: Empirical powers for  $Q$ ,  $Q^{ab}$ ,  $G$ ,  $G^{ab}$  and  $RC^{2/3}$ , for 1 000 samples of size  $n = 400$  and  $R = 100$  categories, at levels  $\alpha = 0.05$  for simulated vector probabilities  $f_1$  to  $f_4$ , and null vector probabilities  $f'_1$  to  $f'_4$ .

$\mathcal{H}_0$	$\mathcal{H}_1$	$mode(c)$	$Q$	$Q^{ab}$	$G$	$G^{ab}$	$RC^{2/3}$
$f_1$	$f'_1$	19	0.233	0.853	0.322	0.983	0.157
$f_2$	$f'_2$	47	0.087	0.026	0.009	0.061	0.009
$f_3$	$f'_3$	64	0.229	0.005	0	0	0.027
$f_4$	$f'_4$	84	0.089	0	0	0	0

Table 4: Diplotype table for the association between TNFAIP3 and Systemic Sclerosis..

Status	H1/H1	H1/H2	H1/H3	H1/H4	H1/H5	H1/H6
Sound	98	7	116	2	71	3
Affected	91	9	104	3	70	12

	H2/H3	H2/H5	H2/H6	H3/H3	H3/H4	H3/H5
Sound	4	2	0	34	1	42
Affected	5	4	1	30	2	40

	H3/H6	H4/H5	H5/H5	H5/H6
Sound	2	1	13	1
Affected	7	1	13	5

hypotheses are written :

$$\mathcal{H}_0 : p_{ij} = p_{i+}p_{+j} \quad \text{against} \quad \mathcal{H}_1 : \exists (i_1, j_1) \in I \times J, p_{i_1 j_1} \neq p_{i_1+}p_{+j_1}, \quad (4.1)$$

where  $p_{i+}$  and  $p_{+j}$  are the marginal distributions for the two characters featured in the table. To ensure the condition (1.7) we remove the empty lines  $i$  of  $\{1, \dots, I\}$  such that  $n_{i+} = 0$ , and the empty columns  $j$  of  $\{1, \dots, J\}$  such that  $n_{+j} = 0$ .

#### 4.1 Multi-marker approach for Systemic Sclerosis

Table 4 is the diplotype table obtained from an association study in Humans looking for an association between three genetic markers on the gene TNFAIP3 and Systemic Sclerosis presented in [8]. Empty columns have been removed. A haplotype is the allelic distribution of markers on a chromosome, and a diplotype is the combination of both parental haplotypes. Diplotype tables, though more interesting than haplotypic tables because they take into account more information, are usually trickier to handle because they are sparse. Our corrected statistics can therefore be helpful in such situations.

Though nine haplotypes theoretically exist, only eight are observed, denoted H1 to H8, leading to  $8^2 = 64$  diplotypes  $H_i/H_j$ . Two samples are compared, affected versus sound subjects, on which we test the independence between the diplotype configuration and the health status of  $n = 794$  individuals.

The table is of dimension  $2 \times 16$ , that is  $R = 32$  categories with  $c = 1$  zero and  $s = 16$  parameters. There are exactly 16 expected frequencies below 5. The chi-square quantile  $\chi_{0.95,15}^2 = 24.99$  is compared to the statistics :

$$Q = 14.62, Q^{ab} = 20.76, G = 15.82, G^{ab} = 28.43, RC^{2/3} = 14.85.$$

Table 5: Contingency table for the joint study of trophic level and vegetable composition in rivers.

Trophic level	$(r, p, e)$					
	(0, 0, 0)	(1, 0, 0)	(0, 1, 0)	(0, 0, 1)	(1, 1, 0)	(0, 1, 1)
Oligotrophic	0	0	3	0	3	2
Mesotrophic	2	1	0	2	1	0
Eutrophic	2	0	3	1	1	0

Only  $G^{ab}$  leads to reject the null hypothesis of independence. This seems to be the right decision since it is confirmed by the single-markers approaches and haplotype tests in [8], all showing a significative association between the markers and the disease.

## 4.2 Trophic level and vegetables in the rivers of the Petite Camargue Alsacienne

The search for a link between the trophic level and the vegetable composition of some rivers of the Petite Camargue Alsacienne in North-East France leads to Table 5. A river can be either oligotrophic, mesotrophic or eutrophic, if its nutritive content is respectively poor, intermediate or high. Uncommon vegetables are considered rare, exotic or polluo-tolerant. To each river, a triplet of binary characteristics  $(r, p, e)$  is assigned, indicating the presence (1) or the absence (0) of rare ( $r$ ), exotic ( $e$ ) and polluo-tolerant ( $p$ ) species. The original ecological study can be found in [11]. We consider  $n = 21$  different rivers. Two empty columns were removed from the original table, leading to Table 5 of dimension  $3 \times 6$ , that is  $R = 18$  categories and  $c = 7$  zeros, with  $s = 17$  parameters. Here are 3 expected frequencies below 0.5.

Test statistics are compared to the chi-square quantile  $\chi_{0.95,10}^2 = 18.31$ :

$$Q = 14.38, Q^{ab} = 20.68, G = 18.67, G^{ab} = 26.05, RC^{2/3} = 14.84.$$

Both corrected statistics  $Q^{ab}$  and  $G^{ab}$  as well as  $G$  lead to reject the null hypothesis, indicating an association between trophic level and vegetable composition. A thorough study of the table shows that rare species tend to settle preferentially in oligotrophic rivers. They are indeed better adapted to this kind of environment which tends to disappear from the rivers in the study. Moreover, polluo-tolerant species constitute the majority of the vegetables in eutrophic rivers. A eutrophic environment is competitive and these resistant species tend to get the best of it.

## 4.3 Discussion

We suggest to compute both  $Q^{ab}$  and  $G^{ab}$ , and to reject the null hypothesis if at least one of them is larger than the chi-square threshold.

Our results tend to prove that this approach is relevant and our corrections efficient in sparse tables. They are all the more interesting for the fact that sparse tables are usually left aside because the hypotheses needed to apply classical chi-square tests are not satisfied.

## A Appendix section

*Proof of Proposition 1.* Assume that the first  $m$  of the  $c$  frequencies  $n_i, 1 \leq i \leq c$  are modified. Let  $n'_j, j \in \{c+1, \dots, R\}$  compensate for these changes. The likelihood inequality (2.2) is then equivalent to :

$$\frac{p_{c+1}^{(n_{c+1}-n'_{c+1})}}{(n_{c+1}-n'_{c+1}+1)!} \times \dots \times \frac{p_R^{(n_R-n'_R)}}{(n_R-n'_R+1)!} \geq \frac{p_1^{n'_1}}{n'_1!} \dots \frac{p_m^{n'_m}}{n'_m!}. \quad (\text{A.1})$$

Let us show that (2.3) implies (A.1). Applying (2.3) to each element of the left hand side of (A.1) with multiplicities such that :  $\sum_{j=c+1}^R (n_j - n'_j) = \sum_{i=1}^m n'_i$ , we get :

$$\frac{p_{c+1}^{(n_{c+1}-n'_{c+1})}}{n^{(n_{c+1}-n'_{c+1}+1)}} \times \dots \times \frac{p_R^{(n_R-n'_R)}}{n^{(n_R-n'_R+1)}} \geq p_1^{n'_1} \dots p_m^{n'_m}. \quad (\text{A.2})$$

As  $(n_j - n'_j + 1)! \leq n^{(n_j - n'_j)}$  for all  $j$  in  $\{c+1, \dots, R\}$  and  $n'_i! \geq 1$  for all  $i$  in  $\{1, \dots, m\}$ , we deduce (A.1) and equivalently (2.2).  $\square$

*Proof of Lemma 2.* Let us first show that:

$$\forall \epsilon > 0, \lim_{n \rightarrow +\infty} \mathbb{P}(C_n > \epsilon) = 0. \quad (\text{A.3})$$

For each  $n$  we compute  $\mathbb{P}(C_n = c)$  for  $c$  in  $\{0, \dots, R-1\}$ . Let  $p^0$  be the probability under the null hypothesis  $p^0 = (p_1^0, \dots, p_c^0, p_{c+1}^0, \dots, p_R^0)$ . A subject belongs to one of the first  $c$  cells with probability  $q^0 = p_1^0 + \dots + p_c^0$  and to one of the  $R-c$  last cells with probability  $1 - q^0 = p_{c+1}^0 + \dots + p_R^0$ , with  $q^0$  in  $]0, 1[$ .

Let us consider now the binomial distribution  $\mathcal{B}(n; q^0)$  and write the probability  $\mathbb{P}(C_n = c)$  of obtaining a table containing exactly  $c$  zeros placed anywhere :

$$\begin{aligned} \mathbb{P}(C_n = c) &= \binom{R}{c} \mathbb{P}(N_1 = \dots = N_c = 0, N_{c+1} \neq 0, \dots, N_R \neq 0) \\ &= \binom{R}{c} (1 - q^0)^n. \end{aligned} \quad (\text{A.5})$$

Then :

$$\mathbb{P}(C_n = 0) = 1 - \sum_{c=1}^{R-1} \mathbb{P}(C_n = c) = 1 - \mathbb{P}(C_n > 0), \quad \forall \epsilon \in ]0, 1[. \quad (\text{A.6})$$

As  $\binom{R}{c}$  is bounded and  $(1 - q^0)^n$  tends to 0, the probability  $\mathbb{P}(C_n = c)$  also converges to 0 for all  $1 \leq c \leq R - 1$  when  $n$  tends to  $+\infty$ , and so does the corresponding sum over  $c$ .

We now use Borel-Cantelli's Lemma to conclude that  $C_n$  converges to 0 almost surely. Indeed, for  $\epsilon > 0$ :

$$\sum_{n \geq 1} \mathbb{P}(C_n > \epsilon) \leq \sum_{n \geq 1} \mathbb{P}(C_n \geq 1) = \sum_{c=1}^{R-1} \binom{R}{c} \frac{1}{q^0} < +\infty. \quad (\text{A.7})$$

□

## Acknowledgements

I want to thank I. Combroux and M. Guedj for letting me use their data sets as illustrations for this work. I am also truly grateful to P. Nobelis for his help and his advice.

## References

- [1] AGRESTI, A. (1990). *Categorical data analysis*. John Wiley & Sons Inc., New York.
- [2] BIRCH, M. W. (1964). A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist.* **35** 817–824.
- [3] BISHOP, Y. M. M. and FIENBERG, S. E. and HOLLAND P. W. (1975). *Discrete multivariate analysis: theory and practice*. The MIT Press, Cambridge, Mass.-London.
- [4] CONOVER, W. J. (1999). *Practical nonparametric statistics*. John Wiley & Sons Inc., New York.
- [5] CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46(3)** 440–464.
- [6] FISHER, R. A. (1922). On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *J. Roy. Stat. Soc.* **85(1)** 87–94.
- [7] GOOD, I. J. (1950). *Probability and the weighing of evidence*. Charles Griffin & Co. Ltd., London.
- [8] GUEDJ, M. *et al.* Association of TNFAIP3 rs5029939 variant with systemic sclerosis in European Caucasian population. *Under review*
- [9] KU, H. H. (1963). A note on contingency tables involving zero frequencies and the  $2\hat{I}$  test. *Technometrics* **5(3)** 398–400.

- [10] PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50** 157–175.
- [11] TRÉMOLIÈRES, M. and COMBROUX, I. and HERMANN, A. and NOBELIS, P. (2007). Conservation status assessment of aquatic habitats within the Rhine floodplain using an index based on macrophytes. *Ann. Limnol.-Int. J. Lim.* **43**(4) 233–244.